Predicting Munich Airbnb Listing Prices

MALIS advanced project done by: Abdessamed QCHOHI, Francesco GIANNUZZO and Alessio GIUFFRIDA

Introduction

- **Problem Statement:**
  - Hosts struggle to set optimal prices.
  - Overpricing reduces occupancy while underpricing leads to lost revenue.
- **Project Goal:**
  - Use machine learning to predict listing prices accurately.
- **Dataset Source:** Inside Airbnb

Related Work

Historical Evolution:

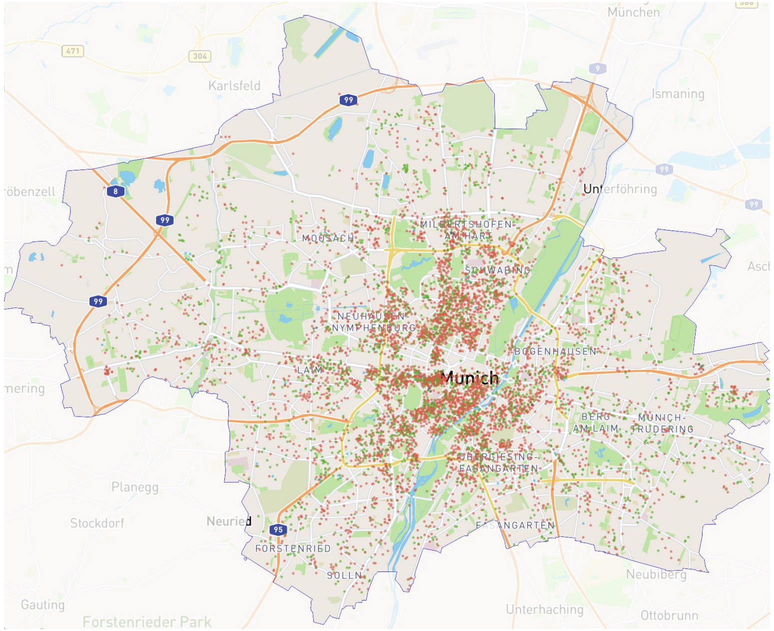
- From linear regression models (Rosen, 1974) to advanced ML techniques.

Key Studies:

- Machine learning models with spatial data improve predictions (Zheng et al., 2019; Yang et al. 2021).
- Relevant previous work applied to Munich rental market (Chen et al.).

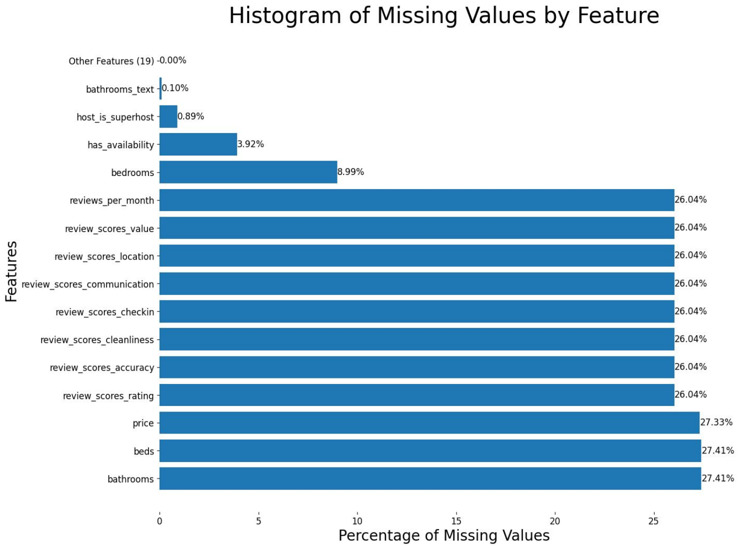
Dataset Overview

- **Dataset Used:** listings.csv.gz
  - 8021 entries, 75 features
  - Many NaN values
- **Feature Selection:**
  - Reduced to 35 relevant features after having removed useless features for the purpose of predicting listings prices
- **Post Data Preprocessing:**
  - Obtained 80 features



Data Preprocessing

- **Handling Missing Values:**
  - Grouped by neighborhood and property_type: mode for price, mean for review_scores.
- **Outlier Detection:**
  - Prices above the 99th percentile removed.
- **Feature Encoding:**
  - Boolean and textual feature transformations.
  - 1-hot encoding after choosing most valuable amenities
  - Normalization of numerical features using StandardScaler



Methods

- **Traditional Regression Models:**
    - Lasso, Ridge Regression (used as baseline).
- **Machine Learning Models:**
    - Random Forest (RF), Gradient Boosting, Stacking Regressor (RF + Gradient Boosting).
- **Deep Learning Approaches:**
    - Multi-Layer Perceptron (MLP), Deep Neural Networks (DNN), DNN with Hypernetworks.

Experimental Setup

- **Train-Test Split:** 80/20
- **Cross-Validation:**
  - K-Fold used except for baseline models and deep neural networks.
- **Evaluation Metrics:**
  - R^2 score and RMSE for performance assessment.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Results Comparison

- **Model Performance:**
  - Best models: Stacking Regressor and Deep Neural Network.
  - Deep Neural Network with Hypernetworks did not significantly outperform traditional models.

- **Numerical results:**
  - Stacking Regressor: $R^2 = 0.7788$ and RMSE = 92.7632
  - Deep Neural Network: $R^2 = 0.7760$ and RMSE = 93.33

Conclusions & Future Work

- No significant improvement with deep learning.
- Limited amount of available data.
- Consider using synthetic data.
- It could be possible to use advanced sentiment analysis.